Volume 3, Issue 11, November 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Rule-based Extraction of Multi-Word Expressions for Elementary Sanskrit Texts

Murali Nandi*
S.V. Vedic Uiversity, Tirupati
India

Ramasree R.J.
Dept. of Computer Science
R.S.Vidyapeetha, Tirupati, India

Abstract: Identification of Multi-Word Expressions (MWEs) in a sentence is essential to understand the sentence beyond semantics i.e., at pragmatic level. So far, not much work was done to extract MWEs. This problem was attempted for Indian Languages by researchers in the form of extraction of chunks of local word groups (LWG). The tools to identify the MWE are language dependent. A tool called Sequence Feature Extractor (SFE) has been developed to extract multi-word groups from elementary Sanskrit texts. In this paper, it is attempted to extract the multi-word groups using morphological analysis (MA) and word sequence rules.

Keywords: Chunks, Local Word Groups, Multi-Word Expressions, Multi-Word Groups, Noun Groups, Phrase, Sanskrit, Sequence Feature Extractor, Verb Groups

I. INTRODUCTION

Sentence (*vaakyam*) means a group of words which expresses a complete idea or thought. The basic element of a sentence is word (*padam*). Each word has its own meaning and expresses a single idea. Some times, the meaning of individual words can not be taken into consideration in the process of analyzing a sentence in order to get the overall meaning of that sentence. This is due to the reason that the phrase (a group of contiguous words) may contribute the meaning other than the combined meaning of the constituent words. For example, the phrase "in front of" in English can not be individually analyzed. The concept of noun phrase and verb phrase cannot be directly applied to Indian languages and hence the term phrase in the context of Indian languages refers to the word groups only [4]. The meaning of a word group or a collection of words generally depends on

- The meaning of individual words and
- The relationship (syntactic and semantic) among words of the group.

However this collective meaning sometimes can not be derived by the individual words of the word group. The reasons for this are given below.

1. Individual words of the group cannot be used independently. (Eg. Case or Vibhakti in almost all languages, as they precede or succeed with nouns)
2. A group of words act like an idiom or phrase, giving separate meaning which is not related to the individual words of the group.
3. For some languages like Hindi, verb and TAM (Tense, Aspect and Modality) are written as separate words; where as in Sanskrit TAM will be indicated by the verb itself in a single word.

In such situations, words need to be grouped together. Generally these word groups may come under two categories viz., noun and verb groups. The next section describes the word groups in brief.

II. WORD GROUPS

The Verb groups and Noun groups can be formed using local or surface information given by the Morphological Analyzer. According to Paaninian Kaaraka theory, these word groups provide sufficient information in processing a sentence.

A. Noun groups

All most all languages have compound nouns. Modifier – Modified relation exists between the words in the group. They also are to be grouped. Noun groups may be formed with the combination of Nouns or Pronouns or Adjectives or *Avyaya* with a noun or pronoun as head word. In Sanskrit the Noun groups may be formed as shown below:

Noun and Nouns → द्वादसे मासे चैत्र-शुक्ल-नवम्याम् पुनर्वसु-नक्षत्र-युक्ते रामम् प्रसूतवती ।
(*Dvaadase maase chaitra-sukla-navamyaam punarvasu-nakshatra-yukte subhe muhuurte kousalyaa raamam prasuutavatii.*)

Nouns and Pronouns → एषः रामः । (Eshah raamah)

→ ऋष्यशृङ्गः एतं प्रस्तावम् अङ्गीकृतवान्।*(Rushyasrungah etam prastaavam angiikrutavaan.)*

Adjective and Noun → सुन्दरः रामः । (Sundarah raamah.)

Nouns and *Avyayas* → तिस्रः अपि गर्भवत्यः जाताः । (Tisrah api garbhavatyah jaawaah.)

B. Verb Groups

Generally in Sanskrit a verb is encoded with the following information.

a. Root of the verb indicating action (धातुः / *dhaatu*)
b. Tense indicating when action occurs (लकारः / *lakaarah*)
c. Gender, Number, Person of the kartaa or karma (लिङ्गम् / *lingam*)
d. Modality (active voice or passive voice) (प्रयोगः / *prayogah*)
e. For whom the action is performed (पदी / *padii*)

In Sanskrit, all the above mentioned information is encoded and written in a single unit (E.g. नमामि). Verb group may be formed as indicated below:

- Verb - Adverb and
- Verb - *Avyaya*

For example,

Adverb and Verb → सः एतं प्रस्तावम् सहर्षम् अङ्गीकृतवान्। (Sah etam prastaavam saharsham angiikrutavaan)

Verb and *Avyaya* → रामः वनं गच्छति स्म। (Raamah vanam gaccati sma.)

Avyaya (*Krit*) and Verb → अहं तया सेनया सः आगत्य युद्धम् कारयिष्यामि । (Aham tayaa senayaa sah aagatya yuxxam kaarayishyaami)

Noun, Pronoun, Adjective, Adverb and Verb are common parts-of-speech. *Avyaya* is something different. The word which cannot be changed or inflected or which remain immutable in all genders, numbers and cases is called as *avyaya*. *Avyaya* i.e., indeclinable words play an important role in the construction of a sentence and can be used as preposition, interjection, particle, conjunction or an adverb. To understand a given sentence, a thorough cognizance of the *avyaya* is necessary. *Avyayas* can be classified into Primitive, *Lakshanika Avyaya*. *Lakshanika Avyayas* can be classified as *Kritpratyayanta* Words as *avyayas*, *Tadditapratyayanta* words as *avyayas* and compounds as *avyayas* [5]. Hence, *avyayas* can occur with verbs as adverbs and they can also appear along with nouns.

In the process of local word grouping, two or more words are to be grouped together into a chunk or a unit and will be treated as a phrase to derive the original meaning of that word group in order to analyze the meaning of the entire sentence. The tool that chunks the local words into a group is called a *Chunker* or *Local Word Grouper*.

III. LOCAL WORD GROUPING OR CHUNKING FOR SANSKRIT

Local word grouping or chunking for Sanskrit is not an easy task. The main problem is ambiguity in lexical category of words, which produces conflict in word grouping. Local Word Grouping has to be done based on the information given by the MA. The MA may give multiple solutions for a particular word and it will be difficult to decide the word's lexical category and its grammatical features like gender, number, person etc.

A. Noun Groups

In Sanskrit, the words which are in same *vibhakti* (case) and *vacana* (number) can be grouped into a unit. The words in the group have a relation called *viseshana – viseshya*. In some cases where more than one person or a thing are to be referred along with a "ca" who/which are participating in the action, then all those words will be in same *vibhakti* and *vacana* and there will not be *viseshana - viseshya* relation. In some cases *avyayas* like *api* should also be grouped into noun groups. The following assumptions have been made to disambiguate the lexical category and other features of a particular word.

- The analysis of verb given by MA is considered to disambiguate other words in the sentence. Based on the analysis of verb, the gender, number and person can be decided for a particular word which is creating ambiguity.
- The word whose lexical category has not been given by the MA or when MA has given multiple parses for a word then, that word can be analyzed based on the adjacent words. If the suffix of the word in question is matching with the suffixes of any one of the adjacent words i.e., previous or next word then the features of the matched words can be applied to the word in question.
- When a *kridanta* is acting as a verb in the sentence, the words in the sentence have been disambiguated based on the *krit* suffix.

B. Verb Groups

A verb group may contain a verb along with particles (*avyayas*) like *sma, khalu, kila* etc. *Avyayas* that are derived from *krit* suffixes will also function as adverbs.

The Local Word Grouper (LWG) or Chunker developed at UoH has been opted for the present study. The Chunker has failed in identifying the word groups in the following cases

- Large sentences consisting of sequence of words in the same *vibhakti* and *vacana*

Example: दृष्टिः जयन्तः जयः सिद्धार्थः अर्थसाधकः अशोकः मन्त्रपालः सुमन्त्रः च इति अष्टौ मन्त्रिणः दशरथस्य ।

Transliteration: *Drishtih Jayantah Jayah Siddharthah Arthasaadhakah Asokah Mantrapaalah Sumantrah ca iti ashtau mantrinah Dasarathasya*.

Meaning: There were eight ministers of *Dasaratha namely Drishti, Jayanta, Jaya, Siddharta, Arthasadhaka, Ashoka, Mantrapala* and *Sumantra*.

- Complex Sentences

Example: यावत् तत् स्थास्यति तावत् भवतः नाम कीर्तिः च अपि जनानाम् रसनासु विलसिष्यति ।

Transliteration: *Yaavat tat sthaasyati taavat bhavatah naama kiirtih ca api janaanaam rasanaasu vilasishyati.*

Meaning: As long as that stands, so long your name and fame will also figure at the tongues of every body.

- *Avyayas* (*ca, iti* and certain adverbs)

Example: सूर्यवंशीयाः राजानः अयोध्याम् राजधानीम् परिकल्प्य कोसलदेशम् पालितवन्तः ।

Transliteration: *Suryavamsiiyaah raajaanah ayodhyaam raajadhaaniim parikalpya kosaladesam paalitavantah*

Meaning: Having made *Ayodhya* as capital, the kings of the dynasty of Sun ruled over *Kosal*.

- Wrong chunking

Example: सः ऐश्वर्येण कुबेरसमानः पराक्रमेण इन्द्रतुल्यः च आसीत् ।

Transliteration: *Sah aisvaryena kuberasamaanah paraakramena indratulyah ca asiit.*

Meaning: He was akin to *Kuber* in opulence and in valor to *Indra*.

To overcome the above mentioned problems a tool called Sequence-Feature Extractor has been developed in PERL and is described below.

IV. SEQUENCE-FEATURE EXTRACTOR (SFE)

The Sequence-Feature Extractor (SFE) deals with the chunking or identifying the multi-words. The input for this module is the output of the MA. The MA gives word-feature pair for each word in the sentence. Based on the information given by the MA and certain rules like *sannidhi* and *akaanksha*, the SFE tries to group the words into noun groups and verb groups. The function of the SFE is described below.

- SFE verifies each word with the previous word and next word. If any one or both the adjacent words are in same *vibhakti* and *vacana*, it will be considered as a group
- SFE verifies that if the given word or word group is in *sashtii vibhakti* or not. If the given word is in *sashtii vibhakti* then, it will be grouped into the noun group which is nearer to it.
- SFE verifies that if the given word is an *avyaya* "*ca*" then, the previous noun groups will be linked to the *avyaya* "*ca*".
- SFE verifies that if the given word is an *avyaya* "*iti*", then the clause present before the *avyaya* "*iti*" will be grouped with "*iti*" and treated as a noun group and will become an object for the verbs like "*avadat*".
- SFE verifies that if the given word is an *avyaya* which functions as an adverb, then it will be grouped with the nearest verb according to the *sannidhi* rule.

The following figure describes the functional model of SFE.

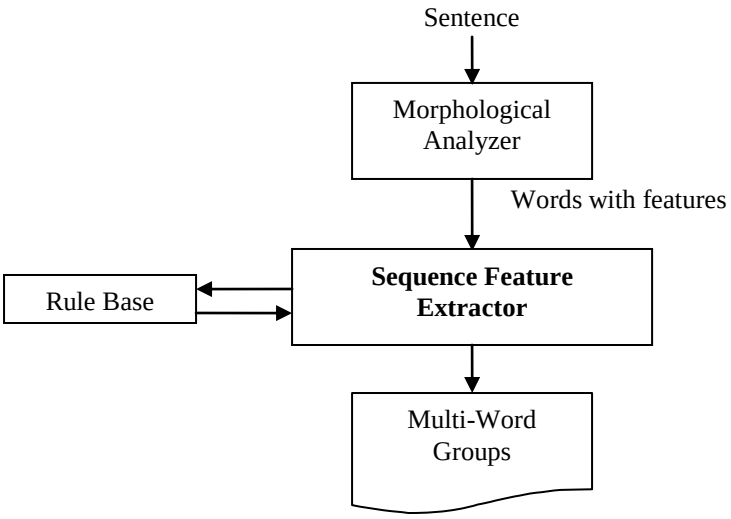

Figure 1: Functional Model of Sequence Feature Extractor

The outputs of the SFE for the sentences given in the previous section are presented below in a table form.

- Large sentences consisting of sequence of words in the same *vibhakti* and *vacana*

Example1: दृष्टिः जयन्तः जयः सिद्धार्थः अर्थसाधकः अशोकः मन्त्रपालः सुमन्त्रः च इति अष्टौ मन्त्रिणः दशरथस्य ।

Transliteration: *Drishtih Jayantah Jayah Siddharthah Arthasaadhakah Asokah Mantrapaalah Sumantrah ca iti ashtau mantrinah Dasarathah*.

Meaning: *There were eight ministers of Dasaratha namely Drishti, Jayanta, Jaya, Siddharta, Arthasadhaka, Ashoka, Mantrapala* and *Sumantra*.

Table I

Words in same *vibhakti* and *vacana* are grouped into a unit by SFE

WNo	Word	Type	Case	No.	Relation	Related	XAwu	Kqw	GaNaH
1	दृष्टिः	ना	1	1	समुच्चितम्	9			
2	जयन्तः	ना	1	1	समुच्चितम्	9			
3	जयः	ना	1	1	समुच्चितम्	9			
4	सिद्धार्थः	ना	1	1	समुच्चितम्	9			
5	अर्थसाधकः	ना	1	1	समुच्चितम्	9			
6	अशोकः	ना	1	1	समुच्चितम्	9			
7	मन्त्रपालः	ना	1	1	समुच्चितम्	9			
8	सुमन्त्रः	ना	1	1	समुच्चितम्	9			
9	च	अव्य	0	0	सम्बन्धः	10			
10	इति	अव्य	0	0	सम्बन्धः	11			
11	अष्टौ	ना	7	1	संख्या	12			
12	मन्त्रिणः	ना	6	1	कर्तृसमानाधिकरणम्	14			
13	दशरथस्य	ना	6	1	षष्ठीसम्बन्धः	12			
14	अस्	क्रि.	0	0	0	0	अस्2		अदादिः

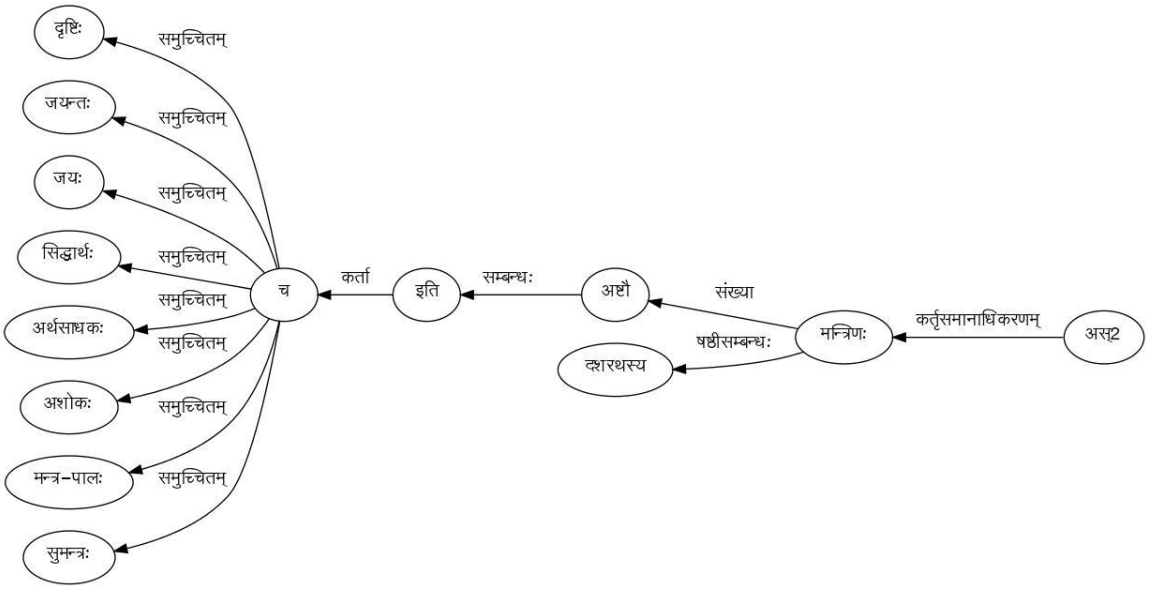

Figure 2: Words in same *vibhakti* and *vacana* are grouped into a unit by SFE

- Complex Sentence

Example 2: यावत् तत् स्थास्यति तावत् भवतः नाम कीर्तिः च अपि जनानाम् रसनासु विलसिष्यति ।

Transliteration: *Yaavaw tat sthaasyati taavat bhavatah naama kiirtih ca api janaanaam rasanaasu vilasishyati.*

Meaning: As long as that stands, so long your name and fame will also figure at the tongues of every body.

Table II

Grouping the words in a complex sentence by SFE

Wno	Word	Type	Case	No.	Relation	Related	Xawu	Kqw	GaNaH
1	यावत्	अव्य	0	0	सम्बन्धः	4			
2	तत्	सर्व	2	1	कर्म	3			
3	स्थास्यति	क्रि.	1	1	प्रतियोगी	1	ष्ठा	0	भ्वादिः
4	तावत्	अव्य	0	0	अनुयोगी	12			

5	भवतः	सर्व	6	2	षष्ठीसम्बन्धः	2	भू	शतृ	भ्वादिः
6	नाम	अव्य	0	0	सम्बन्धः	7			
7	कीर्तिः	ना	1	1	समुच्चितम्	8			
8	च	अव्य	0	0	कर्ता	12			
9	अपि	अव्य	0	0	सम्बन्धः	8			
10	जनानाम्	ना	6	3	षष्ठीसम्बन्धः	11			
11	रसनासु	ना	7	3	अधिकरणम्	12			
12	विलसिष्यति	क्रि.	1	1	क्रि.		लस्	0	भ्वादिः

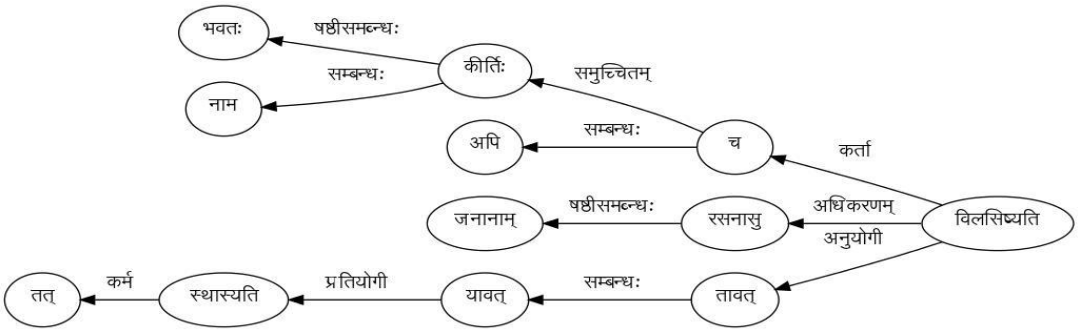

Figure 3: Grouping the words in a complex sentence by SFE

- *Avyayas* (*ca, iti* and certain adverbs)

Example 3: सूर्यवंशीयाः राजानः अयोध्याम् राजधानीम् परिकल्प्य कोसलदेशम् पालितवन्तः ।
Transliteration: *Suryavamsiiyaah raajaanah ayodhyaam raajadhaaniim parikalpya kosaladesam paalitavantah*
Meaning: Having made *Ayodhya* as capital, the kings of the dynasty of Sun ruled over *Kosal*.

Table III
Identification of *Avyaya* as an adverb by SFE

WNo	Word	Type	Case	No.	Relation	Related	XAwu	Kqw	GaNaH
1	सूर्यवंशीयाः	ना	1	3	विशेषणम्	2			
2	राजानः	ना	1	3	कर्ता	7			
3	अयोध्याम्	ना	2	1	विशेषणम्	4			
4	राजधानीम्	ना	2	1	कर्म	5			
5	**परिकल्प्य**	**अव्य**	**0**	**0**	पूर्वकालः	**7**			
6	कोसलदेशम्	ना	2	1	कर्म	7			
7	पालितवन्तः	ना	1	3	0	0	पाल		चुरादिः

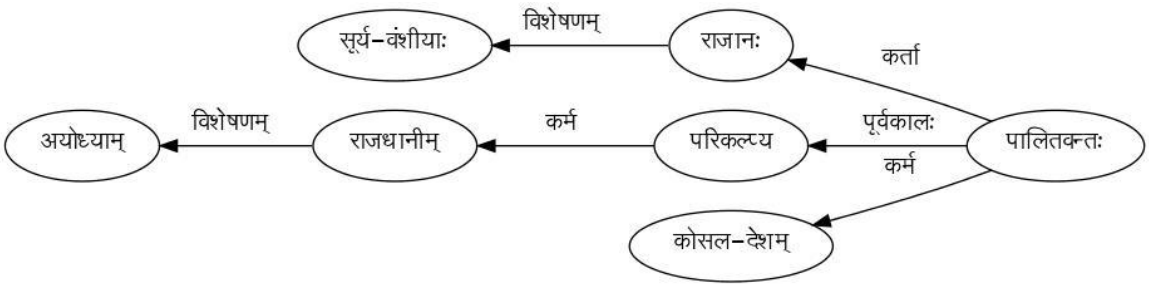

Figure 4: Identification of *Avyaya* as an adverb by SFE

- Wrong chunking

Example 4: सः ऐश्वर्येण कुबेरसमानः पराक्रमेण इन्द्रतुल्यः च आसीत् ।
Transliteration: *Sah aisvaryena kuberasamaanah paraakramena indratulyah ca asiit*.
Meaning: He was akin to *Kuber* in opulence and in valor to *Indra*.

Table IV
Corrected output produced by SFE

WNo	Word	Type	Case	No.	Relation	Related	XAwu	Kqw	GaNaH
1	सः	सर्व	1	1	समुच्चितम्	6			
2	ऐश्वर्येण	ना	3	1	उपपदसंबन्धः	3			
3	कुबेरसमानः	ना	1	1	समुच्चितम्	6			
4	**पराक्रमेण**	**ना**	**3**	**1**	उपपदसंबन्धः	**5**			
5	इन्द्रतुल्यः	ना	1	1	समुच्चितम्	6			
6	च	अव्य	0	0	कर्ता	7			
7	आसीत्	कर्तरि	प्र	1	अभिहित-कर्ता	6			

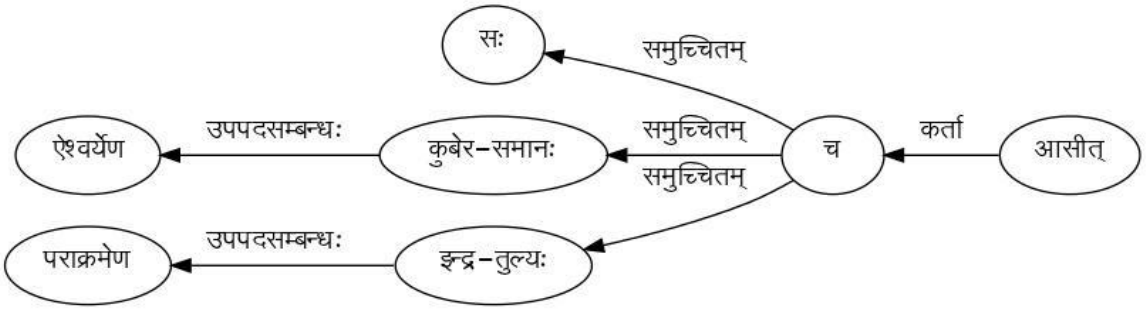

Figure 5: Corrected output produced by SFE

A. Performance issues of Sequence-Feature Extractor

Sanskrit is word free order language. One cannot expect that the words may occur in the same sequence. The word may occur any where in the sentence. But the *vibhakti* and *vacana* helps in maintaining the coherence between the words in the sentence. The SFE has to be improved in this direction. Further, in certain cases the SFE may not be able to identify word groups which are wrongly identified by the Local Word Grouper of the UoH. Such failures have been rectified in the next level i.e., Understanding at Sentence Level.

V. CONCLUSION

Identification of phrases in a sentence is essential to understand the sentence in its correct sense. In Sanskrit, the concept of phrases is implemented through word groups. The word groups can be classified into Noun groups and Verb groups. The process of grouping the adjacent words into a unit is called as Local Word Grouping or Chunking. A tool called Sequence Feature Extractor (SFE) has been developed to optimize the performance of the Chunker developed at UoH. The input for SFE is the word-feature pair given by the (MA). The words which are in same *vibhakti* and *vacana* are grouped into a unit. The features of the adjacent words were applied to the unrecognized word where the suffixes of both words are akin to each other. SFE takes care of conjunctions like "*ca* and *iti* etc."

ACKNOWLEDGEMENTS

We sincerely thank Prof. K.V. Ramakrishnamacharyulu for giving valuable suggestions and Dr. Amba P. Kulkarni who has given the Chunker developed by her at UoH for the present study.

REFERENCES

[1] Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing* (pp. 1-15). Springer Berlin Heidelberg.

[2] Bharati A, Rajeev Sangal, Dipti M Sharma, Sriram V, T Papi Reddy. (2004). Handling Multi-word Expressions without Explicit Linguistic Rules in an MT System. In Proceedings *of Seventh International Conference on TEXT, SPEECH and DIALOGUE - 2004. Brno*, Czeck Republic.

[3] Kunchukuttan, A., & Damani, O. P. (2008). A System for Compound Noun Multiword Expression Extraction for Hindi. In *Proceeding of 6th International Conference on Natural Language Processing (ICON)* (pp. 20-29).

[4] Bharati, A., Chaitanya, V., Sangal, R., & Ramakrishnamacharyulu, K. V. (1995). *Natural language processing: a Paninian perspective* (pp. 65-106). New Delhi: Prentice-Hall of India.

[5] Murali, N., Ramasree, R. J., & Acharyulu, K. V. R. K. (2012). Avyaya Analyzer: Analysis of Indeclinables using Finite State Transducers. *International Journal of Computer Applications*, *38*(6), 7-11.

[6] Birla, V. K., Ahmed, M. N., & Shukla, V. N. Multiword Expression Extraction–Text Processing. *Proceedings of ASCNT-2009, CDAC, Noida, India*, 72-77.

[7] Katz, G., & Giesbrecht, E. (2006, July). Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and*

Exploiting Underlying Properties (pp. 12-19). Association for Computational Linguistics.

[8] Dias, G. (2003, July). Multiword unit hybrid extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18* (pp. 41-48). Association for Computational Linguistics.

[9] Ckakraborty, T. (2010). Identification of Noun-Noun (NN) Collocations as Multi-Word Expressions in Bengali Corpus. In *Student Session, International Conference of NaturalLanguage Processing (ICON)*.

[10] Antoine R. (1968). Sanskrit Manual For High Schools Part – I, 6th Ed. Xavier Publication, Culcutta

